

Will People Like Your Image?

Katharina Schwarz*

Patrick Wieschollek*
University of Tübingen

Hendrik P.A. Lensch

Abstract

The wide distribution of digital devices as well as cheap storage allow us to take series of photos making sure not to miss any specific beautiful moment. Thereby, the huge and constantly growing image assembly makes it quite time-consuming to manually pick the best shots afterwards. Even more challenging, finding the most aesthetically pleasing images that might also be worth sharing is a largely subjective task in which general rules rarely apply. Nowadays, online platforms allow users to “like” or favor certain content with a single click. As we aim to predict the aesthetic quality of images, we now make use of such multi-user agreements. More precisely, we assemble a large data set of 380K images with associated meta information and derive a score to rate how visually pleasing a given photo is. Further, to predict the aesthetic quality of any arbitrary image or video, we transfer the obtained model into a deep learning problem. Our proposed model of aesthetics is validated in a user study. We demonstrate our results on applications for re-sorting photo collections, capturing the best shot on mobile devices and aesthetic key-frame extraction from videos.

1. Introduction

The huge and constantly growing amount of visual online data turned the web into a game. People love to share and rate content what then motivates again to upload even more images and get feedback for free. Thereby, no limits are set. Both, good just like bad images can be found. Besides, many online platforms allow users to “like” or favor certain content with a single click, and, thereby, even further ease the rating process. Usually people tend to “like” beautiful images or, in other words, aesthetically pleasing ones. So, why not take advantage of all this freely and motivated available data? We now aim to explore those multi-user agreements and turn them into a new useful measurement.

Generally, the wide pervasion of digital devices changed the way we deal with media content like photos and videos allowing to take a quick snapshot to make a moment persistent as a photo. As memory storage becomes cheaper there

is no more need to spend time on deleting previously taken images by filtering out mediocre ones. Along with practically no space constraints, we usually do not invest time to organize our photo collections according several criteria. So any great photo you capture today will be useless in one month disappearing in the sheer amount of newly captured photos. Even more, we usually take a series of photos to keep a single moment to not miss the big shot delaying the process of selecting the best shot. Finding particular images in a collection of photos with specific content is not sufficient alone. Instead, we usually want to retrieve the most visually pleasing photos of a specific scene or get a quick overview of the entire collection just depicting the top most aesthetic images.

Recently, deep learning methods shine in categorizing images in discrete sets based on their content. These models can be applied as tagging methods to automatically filter a set of images according to their content and retrieve contextually related images. However, in addition to knowing who or what is in the photo it is likely to ask “how good” is a given image. We are usually only interested in those images which are worth looking at according our sense of aesthetics dropping mediocre photos from the collection.

Object recognition tasks are well defined in terms of the presence of particular objects represented as a binary information. However, rating image aesthetic is highly subjective and fuzzy in terms of classification. We ask the reader to judge the score of image below within the interval $[1, 10]$. While we might have a good intuition on how to



Figure 1. From 1 to 10, how much do you like this image?

score this image based on our experience, it becomes easier if we are asked to compare multiple image selecting the best one. The data-evaluation of 3565 Flickr users reveals that

*Indicates equal contribution.

the judgement task on a single image is highly debatable as the range receives a bi-modal distribution for visually pleasing or mediocre ones. Hence, there is no universally accepted boundary between aesthetic images and undistinguished ones. Publicly available images on the Internet provide some clues indicating how attractive an image is by observing user reactions. We propose to use the observed consensus of multiple users as a proxy to “measure” how visually pleasing a given photo is. As applying data-driven approaches especially on fuzzy tasks requires a large amount of annotations in terms of examples and density, we propose to crawl information from social media platforms such as Flickr. In summary our main contributions are:

- A new large-scale dataset containing dense and diverse meta information such as text information and statistics to reliably predict visual aesthetic.
- Formulating the aesthetic prediction directly as an embedding problem without using labels to directly learn the feature space.
- We show prototypes of applications such as an app for mobile devices, photo-collection manager powered by visual aesthetic prediction as well as a video processing tool to score frames.

2. Related Work

Deep-learning In the last few years deep learning methods such as convolutional neural networks have emerged as powerful image representations for various tasks in object classification [37, 1, 40, 15], object detection [13] and visual question answering [27, 11]. They even allow to manipulate images as demonstrated in [12, 18] by transferring artistic style from a painting to a captured photo. Learning similarities as distances in a feature space allows to arrange images adapted to a specific task. This has been done by training neural networks for signature verification [4], face recognition [7, 42] and comparing image patches [43] for depth estimation. But all these methods rely on datasets with extensive and accurate human annotations as well as reliable ground-truth data for binary decisions.

Aesthetics in Images. Previous approaches on visual aesthetics assessment research mainly focused on handcrafted visual cues such as color [36, 9, 35], texture [9, 20] and content [10, 32].

Regarding general quality of a photograph, e.g., Ke et al. [20] propose a principled method to distinguish between high quality photos and snapshots of low quality. However, judging the aesthetic quality of images has also been previously explored in various ways. Generally, no absolute rules exist to ensure high aesthetic quality of a photograph. Certain heuristics can be applied to improve the pleasingness in terms of composition, e.g. as presented by Jacobitz [17]. Approaches exist that aim at enhancing an existing image

composition [28, 14]. Further, Tong et al. [41] approach aesthetic quality by classifying photos as taken from professional vs laymen. Datta et al. [9] extract visual features based on artistic intuition to differentiate between aesthetically pleasing images and displeasing ones. An overall perspective of aesthetics as well as emotions in images is given by Joshi et al. [19]. Dhar et al. [10] estimate aesthetic quality on attributes humans might use. They select the three broad types composition, content and sky-illumination to train a classifier. Luo et al. [32] use regional and global features to obtain a content-based assessment whereas Lo et al. [29] propose aesthetic features with high computational efficiency. Li et al. [24] restrict their image content and focus on consumer photos showing faces. Similarly, Li et al. [25] develop a photo selection system to manage consumer photos with faces in focus. Besides of photographs, learning computational models of aesthetics has also been considered on other visual domains, e.g., on paintings by Li et al. [23] or evolved abstract images by Campbell [6]. Classifying the aesthetic appeal of consumer videos has been explored by Moorthy et al. [33] and extended by Bhattacharya et al. [3]. Closest to our work are probably Dhar et al. [10], Lu et al. [30] and Kong et al. [21]. Contrarily, instead of applying certain rules or defining specific attributes, we base our method on simple online data that is generated without any challenge or forced design.

Limitations of previous work. These previous approaches underlie several limitations. First, they tend to consider image quality rating as a traditional classification or regression problem predicting a single scalar information real or binary. This shifts the problem from the original ranking nature to a totally different task. Humans probably compare image-pairs rather than measure the aesthetic intrinsically using a scalar. Any reformulation of the original problem introduces several drawbacks. While humans might disagree on the actual level of visual pleasingness they probably rank images in the same aesthetics-order. This suspends any dataset of binary rating-attributes and prohibits the underlying modeling as a classification or regression task. Second, these approaches either use handcrafted features or examine a dataset of small annotation density. As judging visual aesthetics is a highly individual process with different biases it is crucial to include the consensus of a large number of distinct individuals using dense information rather than from a small group of people.

3. Data Sets

As the visual aesthetic of photos is highly subjective depending on the current mood as well as any emotion, training a data-driven model requires extensive annotations. We therefore introduce a new data set in comparison to previous

properties	AVA [34]	AADB [21]	Ours
max ratings*	549	5	2.8M
mean ratings*	210	5	6868
rating distr.	normal	normal	uniform
number of images	250K	10K	380K
avg. image size	602×689	773×955	1926×2344

Table 1. Comparison of different data sets containing images for judging visual pleasingness of images. * Per image

benchmark sets. Table 1 compares our data set to previous ones.

3.1. Previous Data Sets

AVA. As the AVA data set [34] provides 250K images classified in visually well-crafted images and mediocre ones on a fix scale. They are obtained from a professional community of photographic challenges. Through their annotation process only a extremely small amount of annotations are collected in comparison to the dimensions of social network members comprising also non-professional photographers. Note, to reliably judge image aesthetics it is inevitable to consider the consensus of highly diverse participants. Unfortunately, although the AVA data set provides image information including links to the according pages, it is hard to finally access the images.

AADB. Recently, Kong *et al.* [21] introduced a new aesthetics and attributes dataset (AADB) comprising of 10K images. Each individual image score in AADB represents the averaged rating of five AMT (Amazon Mechanical Turk) workers, who are asked to give each image an overall aesthetic score. In addition, they provide attribute assignments from 11 pre-defined categories as judged by AMT workers. Their database maintains photos downloaded directly from Flickr. These are likely to be not edited or post-processed in contrast to professional results contained in AVA [34].

Today's datasets [38, 26] comprises of much larger data enabling training of much deeper networks. We are apprehensive of the small amount of training data given in AADB and biased elicitation of images in AVA. Therefore, we propose to use a new much larger dataset, which can be obtained at almost no time – including labels and meta-data.

3.2. Our Flickr Subset

As a single click allows users to give feedback to media content as images we propose to use this information. E.g., Flickr allows to add any photo to a personal list of favorites, which is counted as “favs”. Since this feature is optionally, users are absolutely free to add a particular image to their favorite list. Their only motivation is to tag a photo which is

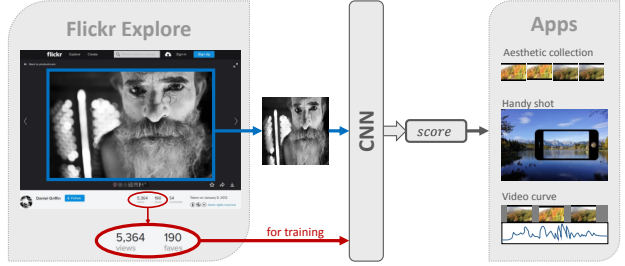


Figure 2. Based on images we assemble from Flickr, an initial score towards image quality is derived from its “views” and “favs” to guide the training process. Our trained CNN is then able to generate predictions for any image leading to several applications.

worth to remember. In addition, these images are uploaded without a purpose to participate in a concrete challenge and are not limited to a specific topic.

To collect these image we crawl around 380K photos from Flickr within 5 days including meta data such as their number of views, comments, favorite list containing this photo, title of the image and their description from the Flickr website. Our collection contains images which were published and uploaded between January 2004 and November 2016. As each photo is visited $\sim 7K$ times in average, this allows for a much finer granularity and gives more hints about aesthetics of images compared to previous data sets.

4. Method

An overview of our method is given in Fig. 2. The main idea is to extract and use time-independent statistics, which contains information traits about the underlying image quality.

4.1. Model of Aesthetics

Previous attempts tried to directly regress some score or trained a simple binary model [35, 9] to decide whether an image is visual pleasing or ordinary. To overcome the classification approaches Kong *et al.* [21] employ a modification of the Siamese loss-function [4] to re-rank image according their predicted aesthetic score.

In contrast to [21, 35, 9], we will leverage traits from freely available information in social networks to score the image quality. These statistics are only used as hints to guide the training process rather than as a direct label or score.

To judge the aesthetic of an image we examine the relation between the “views” (number of visits) and the “favs” (number of clicks that favor image) as a proxy for the visual aesthetics. These both landmarks are highly dependent of the visual aesthetic and encode the visual quality in all it facets. In addition, the low hurdle of creating a feedback (“like” or “fav”) allows to average information being orders of magnitude larger compared to datasets obtained



Figure 3. Some images from our dataset. The upper rows shows images i with large values in $S(i)$ and the bottom row shows examples with relatively low scores $S(i)$.

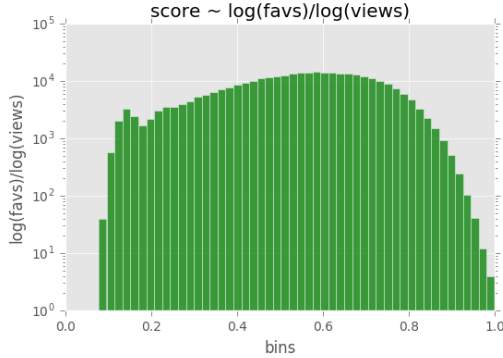


Figure 4. Distribution of the assumed score. Our model assumption leads to uniformly distributed scores.

via AMT. This is especially necessary, when treating images which are highly debatable. As common in population dynamics we assume exponential increase of the views $\frac{dV(i)}{dt} = r_{V(i)} \cdot V(i)$ and the favs $\frac{dF(i)}{dt} = r_{F(i)} \cdot F(i)$ over time $t \in \mathbb{N}$ for any arbitrary image $i \in \mathcal{I}$ with grow rate $r_{(\cdot)} > 0$. This allows us to approximate the score $S(i)$ of the image quality –independent of time t – by

$$S(i) \sim \frac{\log F(i)}{\log V(i)}. \quad (1)$$

This time-independence of any image i is necessary when using images with different online life-spans. In addition, the model in Eq. (1) accounts for the effect of getting more favs per image being a popular user at Flickr due to the mechanism of followers. Note, the action to add an image to ones “favs” contains valuable information, too! Considering the score $S(i)$ gives a criteria to rank images $i \in \mathcal{I}$, which values can be imitated by neural networks, see Figure 3. The distribution of $S(i)$ reveals as mostly uniform iid., which helps to even judge images with border-line score. A histogram of the score distribution from Eq. (1) is illustrated in Figure 4.

4.2. Aesthetic Metric

As the visual quality of images is naturally hard to encode in a single scalar and it is hard to match images to dis-



Figure 5. Image triplets for training with scores $S(i)$. Each triplet consists of either 2 good and 1 bad image concerning its approximated quality or 1 good and 2 bad ones.

crete bins of aesthetic levels, we aim for embedding a given image in a high-dimensional feature space resembling the visual aesthetic in contrast to [21]. This allows to compare images relatively without directly predicting fixed single real or binary value. Ranking approaches like [21] predict scalars and inherently assume that image orders are possible on a single scale. We would like to keep our model flexible enough to consider images, which are similar like consecutive frames in a video. Therefore, we would like to learn high-dimensional embeddings instead of a 1-dimensional ranking space.

Inspired from metric learning by Siamese networks [8] and triplet networks [16], our approach is to indirectly optimize *relative* positions in the embedding space instead of classify images in categories or learn a ranking score. We propose to indirectly learn an aesthetic metric

$$\delta: \mathcal{I} \times \mathcal{I} \rightarrow \mathbb{R}, \quad (i, j) \mapsto \|\Phi_i - \Phi_j\|_2$$

by comparing image pairs (i, j) according to the Euclidean distance of their embeddings Φ_i, Φ_j with unit norm in a learned d -dimensional space.

This allows us to train without any specifically requested ratings or score from human annotators, employing solely the information if two images are similarly aesthetic or not. For any triplet (i, j, k) containing three images $i, j, k \in \mathcal{I}$ we trained a neural network to predict whether an image k is as visual pleasing as i or j . Hence, our training data consists of image triplets

$$D = \{(a, p, n) \text{ with } |S(a) - S(p)| < |S(a) - S(n)|, \\ \text{and } |S(a) - S(p)| < |S(a) - S(n)|\}.$$

An example of such image triples is shown in Fig. 5. Learn the aesthetic metric δ is done by minimizing the triplet loss function

$$L_e(a, p, n;) = \left[m + \|\Phi_a - \Phi_p\|_2^2 - \|\Phi_a - \Phi_n\|_2^2 \right]_+ \quad (2)$$

for images a, p, n and some margin m . Here, $[x]_+$ denotes the non-negative part of x like the ReLU activation function. This loss resembles a visual comparison, i.e., the distance

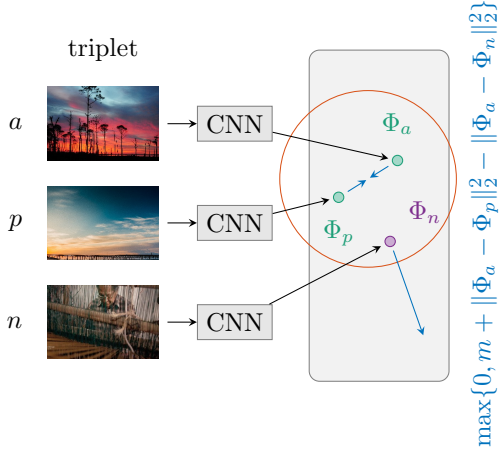


Figure 6. Triplet-Loss. For each image triplet (a, p, n) with anchor point, we aim at embedding aesthetic similar images a, p nearby and force a larger distance to aesthetic dissimilar images n .

between two mediocre images a, p should be smaller than the distance to a well-crafted image n and vice versa. Note that we do not use the actual score $S(\cdot)$ here directly. Instead, we use $S(\cdot)$ to guide random sampling from the data set to find triplets with large enough pair-wise distances. During sampling triplets with mostly indistinguishable distances are rejected to speed up training, i.e., $|S(a) - S(p)|$ should be at most 0.7 times the other distances in the triplet. We allow (a, p) to contain images with higher or lower score than n for generating balance training data. This idea has the following advantages:

1. Every single image can be considered during the training *relatively* to other images, which also allows to train on highly debatable images.
2. There is no need to either learn a scalar or solve a binary classification problem in the fashion of ranking [21] or aesthetic-label prediction [34]. Instead, we learn the embedding itself.

Resolving Direction While the embedding problem by learning distances introduces high flexibility, it hides any ordering of image pairs, i.e. for any two images x, y knowing the aesthetic distance $\|\Phi_x - \Phi_y\|$ has no information if x should be considered as more visually pleasing than y . Therefore, we add a “direction” loss L_d given as

$$L_d(a, n) = \text{sgn}(s(n) - s(a)) \cdot [\|\phi_a\| - \|\phi_n\| + \tilde{m}]_+ \quad (3)$$

$$\Phi_x = \frac{\phi_x}{\|\phi_x\|_2} \quad (4)$$

for $(a, \cdot, n) \in D$ where ϕ_x is the unnormalized embedding of x . This leads the triplet loss by reducing the norms of

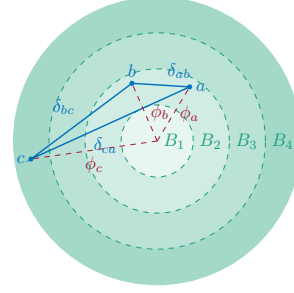


Figure 7. Previous approaches treat aesthetic learning as a low-dimensional problem [21] which projects embeddings on a 1-dimensional or into discrete bins [34]. Rather than learning a bin-mapping for each image $i \in \{a, b, c\}$ into bins B_i or directly ϕ_i , we propose to learn pair-wise distances δ_{ij} .

embeddings belonging to less visual pleasing images and increases the norms of well crafted images. Note, that we again do not use any absolute values from our data model. Altogether, we aim to minimize the “directional triplet loss”:

$$L(a, p, n) = L_e(a, p, n) + L_d(a, n)$$

to get a natural ordering by the euclidean norm and relative distances.

5. Implementation

Throughout this section, we explain the steps in the algorithm in more detail. We use TensorFlow [2] for our prototype implementation running on two Nvidia Titan X GPUs for training.

Model Design We trained a CNN [22] to directly predict the aesthetic similarity by minimizing the directional triplet loss $L(a, p, n)$ introduced in the previous section to adapt the Euclidean distance $\delta(x, y) = \|\Phi_x - \Phi_y\|_2$ between high-dimensional descriptors Φ_i of frames i by learning parameters θ in

$$\Phi_i := f^{(\theta)}(i) = \left(F_n^{(\theta)} \circ F_{n-1}^{(\theta)} \circ \dots \circ F_1^{(\theta)} \right) (i) \quad (5)$$

for network layers $F_k^{(\theta)}$. The ResNet-50 architecture [15] with four blocks b_1, b_2, b_3, b_4 is used for modeling $f^{(\theta)}$ with additional L2 normalization to all parameters. We add a projection from the *pool5* layer creating a 1000-dimensional descriptor for each frame. As CNNs tend to learn edge filters in early layers, we initialize the first Resnet-block using a pretrained version of ResNet [15]. Each parameter θ in the last three ResNet-blocks b_i is updated by stochastic gradient descent with initial learning rate 10^{-3} which is divided by 10 when the error plateaus.

The training progress is verified via cross-validation. Additionally, we constrain this embedding to live on the d -dimensional hypersphere for distance computations, *i.e.* $\|\Phi_i\|_2 = 1$. Note, we optimize the distances of embedding projections on the unit hypersphere in the L_e -loss and embedding norms in the L_d -loss.

Sampling Training Data We randomly sample 250k images for training from our entire collection. Thereby, images without “favs” were rejected. Sampling triplets $(a, p, n) \in D$ is done on-the-fly by randomly choosing three images (a, p, n) and rejecting the entire triplet if it does not match the properties of the set D . We estimate the cardinality of D as $|D| = 7 \cdot 10^{12}$ from tracking the reject-rate during training. Hence, no data-augmentation is required, which would further influence the aesthetic. As ResNet expects the input to have the size $224 \times 224 \times 3$, we resize the original image to match the input dimensions.

6. Evaluation of Aesthetic Model

As we make use of freely available online behavior, the favs and the views from images on Flickr, we want to evaluate our derived metric to demonstrate its usefulness. More precisely, as we introduce our aesthetic model (Sec. 4.1) to approximate the image quality as a score (Eq. (1)) we derive from those uncontrolled user clicks from online media, we validate its sense of purpose in a controlled experiment. We formulate our hypotheses as follows:

1. Our derived “aesthetic model” based on freely available ratings from an uncontrolled human online behavior is reasonable.
2. Higher scored images are also rated better in a controlled user study and worse ones are also rated worse.

Overall, rating the aesthetic quality of an image is highly subjective and differs from person to person. Thus, performing a user study over a broad diversified crowd is inevitable to validate trends. As stated by Buhrmester et al. [5], Amazon Mechanical Turk (AMT) yields reliable data on a demographically diverse level. Therefore, we make use of AMT to evaluate our aesthetic model.

Experiment Setup. To ascertain that images obtaining a higher score are really more pleasant than lower scored ones, we design the study as pairwise preference tests. Thereby, the AMT workers are presented two images with different scores. An example is given in Fig. 8. In each binary forced-choice task, the Turker is asked to select the image that is “aesthetically more pleasing”. We directly ask for aesthetic selection to ensure that our score derived from favoritisation is a suitable measure to rate aesthetics. Overall,

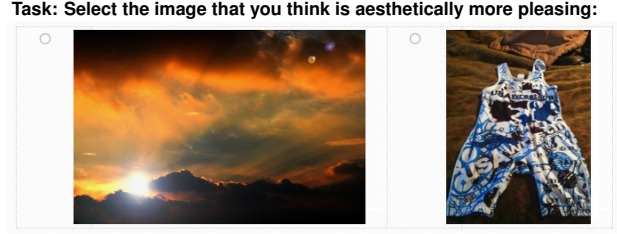


Figure 8. Example as presented on AMT.

from our downloaded data set, we evaluate randomly selected 600 image pairs. Each pair is presented to 5 Turkers. Further, to negate click biases, ordering as well as positioning are randomized.

User Study Results. In our user study, we randomly test image pairs with varying distances between the scores derived by our model. Thereby, the lowest scored images obtained at least one fav. All evaluated distances are listed in Table 2. Thereby, a small distance means that our de-

Table 2. Results of user study. For larger distances $\Delta = |S(i) - S(j)|$ between our derived scores $S(\cdot)$ of the images within a pair, more similar rating decisions of AMT Turkers are obtained.

dist Δ	> 0.1	> 0.2	> 0.3	> 0.4	> 0.5	> 0.6
mean	0.77	0.81	0.84	0.85	0.85	0.89
var	0.21	0.15	0.13	0.13	0.13	0.09

rived scores are very similar and that the images are almost identically pleasing towards aesthetics. However, setting the minimal distance between the scores of the 2 images in a pair to 0.1 is rated towards the similar direction by already 77% of the Turkers. However, considering the biggest score distance we tested in our experiment, namely 0.6, even 89% test persons agreed with the selection of the better image. Besides, the already relatively small variance even constantly decreases with increasing distance. This indicates a high agreement between the different Turkers. Those outcomes of our study demonstrate the correctness of our second hypothesis. As we explicitly ask the Turkers to rate due to the term “aesthetically pleasing”, our presented score $S(i)$ can really be seen as an aesthetic measure and our first formulated hypothesis is also correct.

7. Results

We pursue two ways of evaluation in quantitative evaluation on the common benchmark set and qualitative evaluation to analyze the internal network mechanism. Further results in combination with applications are presented in Section 8 and the supplemental material.

Table 3. Comparison of performance between different models on the AVA dataset. Results indicated by * use additional information during training.

method	accuracy
DCNN [30]	73.25 %
RDCNN* [30]	74.46 %
Reg-Rank+Att* [21]	75.48 %
Reg-Rank+Att+Cont* [21]	77.33 %
Alexnet-FTune [31]	59.09 %
Murray [34]	68.00 %
Reg-Rank [21]	71.50 %
Reg [21]	72.04 %
SPP [30]	72.85 %
DMA [31]	74.46 %
ours	75.83 %

7.1. Quantitative Evaluation

For a fair comparison to previous approaches we fine-tune the last block of our Resnet network on 25K images from the AVA data set and add a linear projection layer. This fine-tuning is necessary, as we have to deal with different distributions of the ratings in the AVA [34] data set compared to ours. Table 3 shows such a quantitative comparison in accuracy to previous methods. Obviously, using an indirect approach such as ranking (Reg+Rank [21]), which resemble the nature of aesthetic judgments much better than standard approaches like classification [30, 31, 34] yields also better performance on this benchmark set. Ours further boosts this accuracy. Results indicated by * use additional information such as attribute data or content-description. Hence, although we trained on a dataset which was constructed with literally no extensive explicit labeling, we outperform all previous methods relying solely on ratings or scores, which are obtained in an expensive process. Further, learning from the consensus of many Flickr users is sufficient to gain higher accuracy (our network) on the AVA benchmark set than recent approaches with additional attributes (Reg+Rank+Attr, RDCNN). Note, these attribute categories acting essentially as a prior and were selected after consulting professional photographers [21].

We expect to further improve our results when adding more explicit information about the content like in the construction of “Reg+Rank+Att+Cont”. As our main focus is to exploit freely available information solely, these explicit meta-information can be image-related comments and tags.

7.2. Qualitative Evaluation

What is the network looking for? Judging the visual quality of an image is totally different from plain object recognition tasks. When extracting relevant information, which is used by the neural network to perform aesthetics

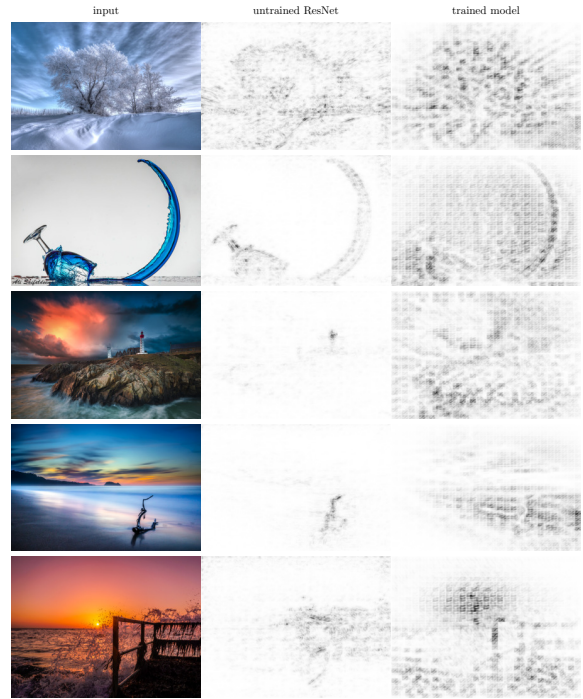


Figure 9. Different photographs (left) and related saliency maps for vanilla Resnet (middle) and our model (right) produced by guided-ReLU [39]. Darker region indicates higher influence on the actual network prediction.

prediction, it is possible to visualize prominent traits in the input. To extract these saliency maps, we use guided-ReLU [39]. It is based on the idea, that large gradients of the output wrt. to the input have a high impact on the actual network prediction. Fig. 9 highlights those pixels in the input with large impact. Hence, this information is strongly coupled with the embedding in our aesthetic feature-space. It clearly shows how our network considers larger regions in the image space compared to sparse saliency along gradients in the untrained network. More precisely, the network model reveals high synergy effects between surrounding regions in the light-house example in Fig. 9. At same time the vanilla ResNet only focuses on the light-house itself.

8. Applications

In order to demonstrate the usability of our approach, we apply our derived aesthetics prediction score to images as well as videos allowing for several applications.

Aesthetic Photo Collection. First of all, we support resorting an arbitrary photo collection due to our predicted relative aesthetic scores between the images. An example of a small set of aesthetically sorted images is shown in Fig. 10. This tool can facilitate to quickly resort one’s holi-



Figure 10. Aesthetically resorted set of photos with decreasing score from our provided tool starting with the visually most pleasing image (left).



Figure 11. Best handy shot. Based on slight movements in any direction, the application automatically captures the best shot.



Figure 12. Best predicted image (red frame) during capturing. The movements were recorded with a mobile device.

day collection and directly share the best moments without time-consuming manually browsing of the usually rather large set of pictures.

Best Handy Shot. A commonly known situation is that people want to take a picture but are not completely sure what could be the best shot of the view. In such situations, they tend to take multiple pictures and just postpone the decision process. This can even lead to missing the one best shot completely. Therefore, we provide a simple application that allows to slightly move the phone around and temporarily captures a video. The idea is visualized in Fig. 11. All the single images are then analyzed and rated by our system and the image of the best view is saved. The application supports the user to directly obtain the best aesthetically pleasing image and prevents the time-consuming decision process afterwards. Fig. 12 shows several frames from movements we recorded with a Samsung Galaxy SII phone and the predicted best shots. Thereby, sky proportions, saturation and the tension of the overall image layout

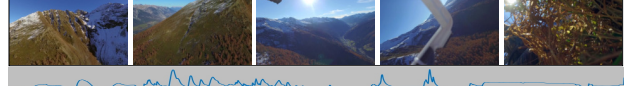


Figure 13. Best video spots. Each frame is extracted at the peaks in the score signal.

play an important role within the decision. Due to its small memory footprint of only 102MB containing the network weights, running this application directly on mobile devices is easily possible. Please see the supplemental video for a short demo. This application could further be extended to lead the user to the best shot during the movement while indicating better directions.

Video Spots. Similarly, our system is able to find great shots in a video. Those shots can be selected as aesthetic key frames or, for example, in documentary films, to identify the most wonderful places or spots. Therefore, we calculate a complete prediction curve along the video displaying the aesthetic relation between the frames. Fig. 13 displays an example of a video and the according aesthetic prediction curve. Kalman filtering is applied to smooth the final predictions over time. Extracting the frame scores is done at a speed of 140fps on a NVidia GTX960. Hence, embedding common videos requires only 25% of the actual playback time demonstrating high efficiency and enabling real-time applications. Please see the supplemental material for example videos.

9. Conclusion

We present a new data set with 380K images and, thus, much larger than previous benchmark sets. The data set was obtained with literally no labeling-costs utilizing freely available information voluntarily given by a huge amount of users. Our data set can be easily extended to more images as needed. For future work, meta data associated with all the images could further be exploited. Formulating the aesthetic prediction directly as an embedding problem with only using implicit labels indirectly, our network learns a feature space to represent the visual aesthetic. Our used directional triplet loss naturally resembles the relative aesthetics judgment of humans. Our model outperforms all previous methods, which do not use additional meta data, on the AVA benchmark set. Hereby, we trained our model on much weaker information. Aesthetics ground-truth scores for training are constructed without explicitly collected user feedback solely for this purpose. The assumption of our underlying model was validated in a small user study. Finally, we demonstrate the success of our model in several applications, namely, resorting photo collections, capturing the best shot and a smooth prediction along a video stream.

References

- [1] Imagenet classification with deep convolutional neural networks. **2**
- [2] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. **5**
- [3] S. Bhattacharya, B. Nojavanasghari, D. Liu, T. Chen, S.-F. Chang, and M. Shah. Towards a Comprehensive Computational Model for Aesthetic Assessment of Videos. In *ACM Multimedia*, Grand Challenge, October 2013. **2**
- [4] J. Bromley, I. Guyon, Y. Lecun, E. Sckinger, and R. Shah. Signature verification using a "siamese" time delay neural network. 1994. **2, 3**
- [5] M. Buhrmester, T. Kwang, and S. Gosling. Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5, 2011. **6**
- [6] A. Campbell, V. Ciesielski, and A. K. Qin. *Feature Discovery by Deep Learning for Aesthetic Analysis of Evolved Abstract Images*, pages 27–38. Springer International Publishing, 2015. **2**
- [7] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. pages 539–546. IEEE Computer Society, 2005. **2**
- [8] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 539–546, Washington, DC, USA, 2005. IEEE Computer Society. **4**
- [9] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, pages 288–301. Springer-Verlag, 2006. **2, 3**
- [10] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *CVPR*, pages 1657–1664, 2011. **2**
- [11] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In J. Su, X. Carreras, and K. Duh, editors, *EMNLP*, pages 457–468. The Association for Computational Linguistics, 2016. **2**
- [12] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015. **2**
- [13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. **2**
- [14] Y. Guo, M. Liu, T. Gu, and W. Wang. Improving Photo Composition Elegantly: Considering Image Similarity During Composition. *Comput. Graph. Forum*, 2012. **2**
- [15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. **2, 5**
- [16] E. Hoffer and N. Ailon. Deep metric learning using triplet network. *ICLR*, 2015. **4**
- [17] S. Jacobitz. <http://petapixel.com/2016/08/08/understanding-basic-aesthetics-photography/>. Accessed: 2016-10-27. **2**
- [18] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016. **2**
- [19] D. Joshi, R. Datta, E. A. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo. Aesthetics and Emotions in Images. *IEEE Signal Process. Mag.*, 28(5):94–115, 2011. **2**
- [20] Y. Ke, X. Tang, and F. Jing. The Design of High-Level Features for Photo Quality Assessment. *CVPR*, 01(undefined):419–426, 2006. **2**
- [21] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision (ECCV)*, 2016. **2, 3, 4, 5, 7**
- [22] Y. LeCun, L. Jackel, L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. Muller, E. Sackinger, P. Simard, et al. Learning algorithms for classification: A comparison on handwritten digit recognition. *Neural networks: the statistical mechanics perspective*, 261:276, 1995. **5**
- [23] C. Li and T. Chen. Aesthetic visual quality assessment of paintings. *IEEE Journal of Selected Topics in Signal Processing*, 3(2):236–252, 2009. **2**
- [24] C. Li, A. Gallagher, A. C. Loui, and T. Chen. Aesthetic quality assessment of consumer photos with faces. In *2010 IEEE International Conference on Image Processing*, pages 3221–3224, Sept 2010. **2**
- [25] C. Li, A. C. Loui, and T. Chen. Towards aesthetics: A photo quality assessment and photo selection system. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 827–830. ACM, 2010. **2**
- [26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, Zrich, 2014. **3**
- [27] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *International Conference on Computer Vision (ICCV)*, 2015. **2**
- [28] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or. Optimizing Photo Composition. *Comput. Graph. Forum*, 29(2):469–478, 2010. **2**
- [29] K. Lo, K. Liu, and C. Chen. Assessment of photo aesthetics with efficiency. In *ICPR*, pages 2186–2189. IEEE Computer Society, 2012. **2**
- [30] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. RAPID: Rating Pictorial Aesthetics Using Deep Learning. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM, pages 457–466. ACM, 2014. **2, 7**

- [31] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *ICCV*, pages 990–998. IEEE Computer Society, 2015. [7](#)
- [32] W. Luo, X. Wang, and X. Tang. Content-based photo quality assessment. In *ICCV*, pages 2206–2213, 2011. [2](#)
- [33] A. K. Moorthy, P. Obrador, and N. Oliver. *Towards Computational Models of the Visual Aesthetic Appeal of Consumer Videos*, pages 1–14. Springer Berlin Heidelberg, 2010. [2](#)
- [34] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2408–2415. IEEE, 2012. [3](#), [5](#), [7](#)
- [35] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato. Aesthetic quality classification of photographs based on color harmony. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '11*, pages 33–40, Washington, DC, USA, 2011. IEEE Computer Society. [2](#), [3](#)
- [36] P. O'Donovan, A. Agarwala, and A. Hertzmann. Color compatibility from large datasets. *ACM Trans. Graph.*, 30(4):63, 2011. [2](#)
- [37] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014. [2](#)
- [38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. [3](#)
- [39] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *arXiv:1412.6806, also appeared at ICLR 2015 Workshop Track*, 2015. [7](#)
- [40] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. [2](#)
- [41] H. Tong, M. Li, H.-J. Zhang, J. He, and C. Zhang. Classification of digital photos taken by photographers or home users. In *PCM*, pages 198–205, 2004. [2](#)
- [42] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. *CoRR*, abs/1607.08378, 2016. [2](#)
- [43] S. Zagoruyko and N. Komodakis. Learning to compare image patches via convolutional neural networks. June 2015. [2](#)